Mining Textual Significant Expressions Reflecting Opinions in Natural Languages

Jan Žižka and František Dařena
Department of Informatics / SoNet Research Center
Mendel University in Brno
Brno, Czech Republic
jan.zizka@mendelu.cz, frantisek.darena@mendelu.cz

***Abstract*—Revealing an opinion hidden in a text document is a challenging task. The article presents a method based on the automatic extraction of expressions that are significant for specifying a document attitude to a given topic. The significant expressions are composed using revealed significant words in the documents. The significant words are selected by the *c5* decision-tree generator based on the entropy minimization. Words included in branches represent kernels of the significant expressions. The full expressions are composed of the significant words and words surrounding them in the original documents. Such expressions provide much more information than individual (key-)words and can be used for analysing a document meaning and the cause of the opinion: what exactly the opinion deals with? The results are demonstrated using large real-world multilingual data representing customers' opinions written in a free form.**

***Keywords*-text mining; natural languages; opinion analysis; significant words; significant expressions; machine learning; decision tree;**

I. Introduction

One of possible text-mining applications includes revealing important, or better to say "kernel" units of text documents that represent the document meanings from the semantic point of view. This meaning is hidden in certain words, phrases, sentences, paragraphs, and so like, however, not all such linguistic items represent the meaning itself. Ordinarily, key-words can present a kind of a document kernel, with all positive and negative properties of this text document representation – this is generally well known, for example, from using Internet web-browser functions "search-for". Individual key-words do not often bring information enough despite the fact they are significant for a given topic. More information is covered by collocations constructed from several significant words – usually, such the collocations are built from some nouns which can better narrow down the searching process. Still, this filtering process is not always quite satisfactory (see e.g. [13]).

Let us consider a more complicated case: The main topic of a collection of text documents written in a certain natural language can be the same, however, the documents contain different attitudes to the topic. This attitude is expressed as an opinion (sentiment). Individual key-words are not enough because, for example, we can say *good* or *bad* but it can also be *not good* or *not bad*, depending on the particular word sequence that, in addition, can be separated by other words: *It is not, in any case, a really very good idea*. The meaning is evidently negative, however, how a machine can recognize it when there is just one "negative" word *not* which is not a very close neighbor to *good*? This situation is typical for any natural language; still, because of very large textual data, people need to automatically process it.

A possible solution can be based on looking for "typical" positive and negative phrases (grammatically correct word sequences). Unfortunately, the extraction of syntactic phrases requires too complex language processing [3] and classifiers using syntactic phrases usually perform worse than classifiers based on the much more simple *BoW* (bag-of-words) procedure [7] where text documents are disassembled to individual words having no mutual relationship [12].

Experienced people doing bibliographic search can reveal the significant words and particularly phrases for languages they are using, however, for millions of text documents and tens of languages the situation is very difficult, needing a certain help of machines. Therefore, the question is: For a given topic, can a computer reveal *phrases* that express a certain opinion, without the exacting and time consuming linguistic analysis which is miscellaneous for different natural languages?

In the following sections, the authors present results of a research work aimed at the mentioned problem. For a very large real-world data set in more than 20 languages, the research applied a decision-tree generator to founding words, here called *significant words*, that reflect opinions hidden in text documents. *Significant expressions* are here defined as whatever sequences of any words included in reviews, without any particular ranking of words, while a sequence has to contain at least one significant word. The significant expressions can be composed from any word classes (nouns, verbs, adjectives, and others) and no grammatical rules are requested – this simplified approach enables avoiding specific language grammars; the price is naturally a certain information loss.

II. Material and Methods

A. Data sources and processing

The research was based on the real-world data containing customers' opinions of hotel accommodations booked via an

978-1-4577-1676-8/11/$26.00

Internet service. The on-line hotel-reservation web-service provides information on hotel prices, facilities, policies, terms, and conditions, including a possibility to type in (via a computer) users' reviews related to their stay in a given hotel. The reviews could not be entered by whichever person but only by the people that made a reservation through the web and stayed in the hotel. Each review consisted of a reviewer identification, his or her overall evaluation (a number on a 10-point scale) and a free review text itself in a selected language. Typically the reviews contained two components, negative and positive experience with the hotel. Because of the policies of the hotel reservation web, the samples were labeled as positive and negative quite carefully.

Using a whatever natural language and script in the unstructured free form introduces many typical problems (mistypings, syntactic and semantic errors, applying various original interjections, transposed and missing letters, sometimes combinations of two languages within one review, and so like). In addition, customers using languages as Czech, Polish, German, Spanish, Russian, French, and others, sometimes apply diacritic, sometimes do not (*nepříjemný/neprijemny*, *möglich/moeglich/moglich*, *mañana/mana*, *maître/maitre)*, which increases the high number of "unique" words (that is, the vector space dimensionality). Additional complication comes from using different alphabets: Chinese, Hebrew, Japanese (using Kanji, Hiragana, and Katakana), Korean, Russian, Serbian, Thai, and others, see for example [9]. Due to the limited space, the authors publish here the results only for the most used languages in the processed data, plus their native one: Czech (ca 17,000 customer reviews), English (1,919,000), German (512,000), and Spanish (470,000). However, the results for other languages were very similar. For each individual language, its reviews were transformed to – typically very sparse – vectors using the well known common pre-processing procedures, including the bag-of-words (BoW) method [7] where the individual words were represented by their frequencies in individual reviews (no numbers or special characters were used; just lower-case alphabetic characters). The simply created bag-of-words suffered from obvious, commonly known shortages, for example, containing several variants of the same word – the experiments, however, had no available batch-mode stemming tools for most of all languages. Therefore, the authors decided to use all word forms because here the goal of experiments was not to reach the best possible classification, which is often an intention. The preprocessing also did not use removing *stop-words* because up to now, due to many different languages and providing the same conditions to each of them, there was not time enough to create lists of such words – this is a task for the near future. Each vector was then labeled according to its class (either the *positive* or *negative* opinion). All the reviews together created the dictionary.

B. Extraction of significant words using a decision tree

A *significant word* is an attribute relevant to labeling a review according its opinion category. Therefore, the first step includes discovering the significant words.

The significant words as the relevant attributes were looked for using decision trees generated from training data which were the available reviews. Each tree branch leads from the root with the most significant word to a leaf that represents the opinion category – in reality, such a branch is the source of significant expressions because it contains significant words (see Fig. 1). Words that are not significant (from the opinion categorization viewpoint) do not occur in the tree. The authors applied the commercially available decision-tree generator *c5/See5* [10] based on the entropy minimization. In principle, the *c5/See5* algorithm, coming from the well-known and popular algorithms *ID3* and *c4.5* [11], splits a heterogenous set (having a non-zero entropy value due to containing items belonging to different classes) into more homogeneous subsets (with most or all items from only one class, that is, with lower entropy than the original set). Recursively, the algorithm tests successively all attributes (here words) and selects only those ones that decrease the entropy. Ideally, all lists should be perfectly homogeneous, containing as many one-class items as possible; in reality, the classification accuracy is often less than 100%.

The average disorderedness, $\bar{H}$, measured by entropy, is given by the disorderedness contributions of nodes at the ends of branches leading off their parent-node which inquires values of a given attribute [11]:

$$\bar{H} = \left[\sum_b \frac{n_b}{n_t}\right] \cdot \left[-\sum_c \frac{n_{bc}}{n_b} \cdot \log \frac{n_{bc}}{n_b}\right], \qquad (1)$$

where n_b is the number of samples covered by a branch b, n_t is the total number of samples in all branches, and n_{bc} is the total number of samples in a branch b for a class c. The disorderedness of a subset for each branch is weighted by the subset size (the number of samples in a subset) relatively to the total size of subsets in all branches, n_b/n_t. The entropy, H, is calculated using probabilities: $H = -\sum_n p_n \cdot log(p_n)$, where, in our case, $p_n = n_{bc}/n_b$, see for example [2]. In the experiments carried out by this research, the attributes were words represented numerically by their frequencies in reviews. Note the fact that no mutual dependence between attributes is assumed – the *BoW* procedure result meets this assumption, too.

From the categorization viewpoint, the real significance of words is related to the classification accuracy – in the described cases, the error was between 7.5-10.5% (typically lower for larger data and vice versa). It is necessary to emphasize that the goal was *not* to reach as good classification accuracy as possible – the goal was to *find relevant attributes* (the significance of which is naturally directly proportional

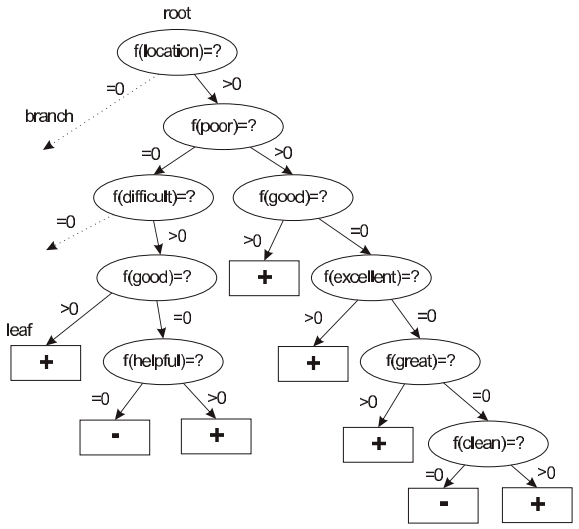

Figure 1. The illustration of a decision tree fragment. A node asks what is the frequency of a word, f(word), in a review. Each branch ends either in a $+$ or $-$ leaf (the positive or negative opinion) according to words in a review. In the tree, only significant words are selected depending on their ability to decrease entropy.

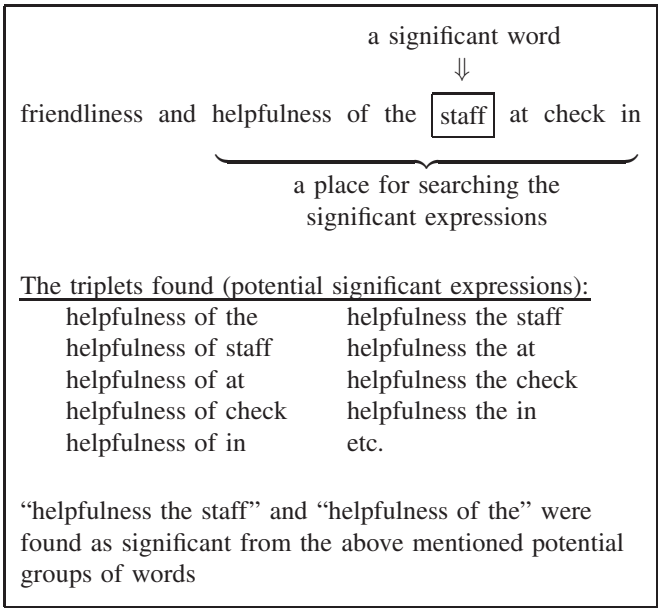

Figure 2. An example of searching for significant expressions in a review (the algorithm parameters: D = 3, N = 3).

to the final accuracy). This approach noticeably decreased the number of all words (in the order of 10^4 to 10^5) to the number of 200–300 significant words. The decision-tree algorithm was chosen because it represented knowledge that was transparent and understandable for the following processing aimed at looking for significant expressions included in the original texts of the customers' reviews. For more details on extracting of significant words and on determining the strength of their positivity and negativity, see [14] and [4].

C. Extraction of significant expressions using a window

The extraction of significant expressions starts from the list of significant words that were identified in the previous step in order to do the process computationally feasible (because of too many possible combinations of words). The significant expressions are therefore searched for not in the entire text but only in the proximity of words significant for classification according the sentiment embodied in the text.

For the significant-expression extracting algorithm, several parameters needed to be determined:

- D – the distance from a significant word within which the search will be carried out. This distance is measured in the number of words.
- N – the number of words forming the significant expressions.
- M – the minimal number of occurrences of a specific group of words so it can be considered as a significant expression. The more times the same group of words appears, the more likely it becomes a significant expression.

During the extraction of significant expressions, all possible combinations consisting of N words within the distance D (to the left and right) from each significant word in a given text were created. Within each of the texts, all possible significant words were searched for (see Fig. 2). For each combination that was found, the number of its occurrences (O_i) was increased by 1. The order of words in the combination was not important, so for example triplets as "very good location" and "location very good" were considered to be the same significant expression.

After processing of all texts, only the combinations with the number of occurrences $O_i \geq M$ were interpreted as significant expressions and added to the output. The significant expressions were also ordered according their frequency so the more frequent ones appeared at the beginning of the list. For just several significant words, several hundreds of thousands of potential word combinations were often found and needed to be considered, the computational complexity was thus very high.

Because the order of individual words within the significant expressions was not preserved, most of the significant expressions were not linguistically correct. However, because the number of words varied from two to five, they were still comprehensible for humans.

III. Results

As a result, the list of significant expressions extracted from the original text reviews was obtained. The quality of these expressions is difficult to measure automatically and

Table I
SIGNIFICANT EXPRESSIONS IN ENGLISH

Positive	Negative
Original	**Original**
and close to the airport	the bathroom was not clean
close to the city center	the room was little
room was comfortable clean	the breakfast was very poor
the room comfortable and spacious	the room was not clean
and the service was excellent	I did not like the
was good value for money	in front of the hotel
very friendly and very helpful	we've got a smoking room
I would stay there again	and was not drinkable tap water
was very kind and helpful	extra for internet for breakfast
and the staff were professional	the bed was very uncomfortable

Table II
SIGNIFICANT EXPRESSIONS IN GERMAN

Positive	
Original	**Translation**
und das Zimmer war gross	and the room was big
Personal sehr freundlich und zuvorkommend	staff very friendly and helpful
man kann alles zu Fuss erreichen	one can reach everything by walk
das Zimmer und bad sehr sauber	the room and bathroom very clean
ein Paar Schritten vom Strand Weg	a few steps from beach path
Terrase mit Blick auf das Meer	terrase with view of the sea
sehr gute Lage ansprechendes Design	very good location attractive design
die Dusche war gross und schön	shower was big and nice
das Hotel is zentral gelegen	hotel is centrally located
das Frühstück war sehr gut	breakfast was very good

Negative	
Original	**Translation**
die Bar und das Restaurant geschlossen	the bar and the restaurant closed
Duschen und Waschbecken sind sehr klein	shower and basin are very small
die Klimaanlage nicht funktioniert	air-contitioning not working
Zimmer extrem klein für den Preis	extremely small room for the price
die Sterne Hotels nicht entsprechen	stars of the hotel don't correspond
so das war nicht schön Zimmer	so it was not a nice room
ist sehr schwer finden das Hotel	is difficult to find the hotel
die Fenster waren nicht sauber	windows were not clean
Zimmer zur Strasse sind sehr laut	rooms to the street are noisy
für Leute mit Rückenleiden Betten nich	no beds for people with back problems

therefore it must be considered, for now, by humans. The significant expressions were mostly quite meaningful (with the order of words which usually does not correspond to the order used in natural language) and potentially useful for the target audience (for example, a hotel manager can see that people mostly appreciate a good location, sea view, friendly staff, et cetera, and are not satisfied with expensive breakfasts, small rooms, difficult parking, noise, and so like). Good examples of the positive and negative significant expressions extracted from texts written in English, German, Spanish, and Czech are demonstrated in Tables I, II, III and IV. If necessary, the order of words in these expressions was adjusted to make them better readable for humans.

Of course, some of the significant expressions were very similar (they differed, for example, in just one word), but ordering them according the number of occurrences in the texts and putting the most frequent ones to the top of the list might contribute to solving this problem. Also excluding the significant expressions with more than, for example, half of the stop-words might filter out ineffective phrases. Without such filtering, the result contains "funny" significant expressions containing obviously important words but missing other words bearing the sentiment, like "and for the breakfast", "and on the to floor", "an der rezepzion war sehr" ("on the reception was very"), "die des des und zimmers" ('the the the and room"), and so like, as well as significant expressions containing neither important words nor words bearing the sentiment, like "by I of the was", "had a I and to", "lo de es que" ("what of is that"), "aber der es in nur" ("but the it in only"), "había de no y" ("and had no"), and similar ones.

IV. DISCUSSION

Allowing the algorithm to search for significant expressions in larger pieces of the text has two antagonistic and conflicting effects. The positive effects include:

- The extracted expressions could be longer (containing more words), contain more information and thus have higher value (for example, "staff was friendly and helpful" is a better information that just "staff was helpful" or "staff was friendly" or "helpful and friendly").
- The really important groups of words from the text can be more likely discovered, the significant expressions do not mostly contain words without significant meaning (stop-words).
- The significant expressions can be found even if the words forming the expressions are not too close to each other, see the examples in Table V.

On the other hand, allowing the search to be performed on longer pieces of texts has also several negative effects:

- The likelihood that the significant words (that will later form one significant expression) will be searched within two or more different contexts is higher. It can be clearly seen in the examples from the texts "... breakfast was really good. The location is a little out of the center ..." or "Good service. Convenient location". Here the words "good" and "location" are two, respectively three words apart (this distance can be, of course, much longer in different texts) but they are obviously not related together. The search could provide better results when it is carried out only within individual

Table III
SIGNIFICANT EXPRESSIONS IN SPANISH

Positive	
Original	**Translation**
la amabilidad del personal	friendliness of staff
servicio de transporte el aeropuerto	airport transportation service
habitación muy amplia y cómoda	room very spacious and comfortable
la tranquilidad del entorno	quiet environment
esta cerca de la playa	is close to the beach
la relación calidad precio excelente	excelent price/quality relation
la atención del personal del hotel	attention of the hotel staff
la terraza vistas al mar	terrace with sea view
muy limpio y todo nuevo	very nice and all new
cerca de la estación tren	close to the train station

Negative	
Original	**Translation**
no tiene ascensor	doesn't have a lift
el desayuno un poco pobre	breakfast a bit poor
y el parquing es dificil	parking is difficult
la falta de aire acondicionado	missing air conditioning
las paredes son de papel	walls are from paper
no me ha gustado	I did not like
las habitaciones son pequeñas	rooms are small
ruido de la calle	street noise
deja mucho que desear	falls short of our expectations
relación calidad precio es mala	price/quality relation is bad

Table IV
SIGNIFICANT EXPRESSIONS IN CZECH

Positive	
Original	**Translation**
vše bylo v naprostém pořádku	everything was absolutely all right
líbil se mi přístup personálu	I liked the behavior of staff
možnost parkovat v blízkosti hotelu	possibility of parking near the hotel
byli jsme spokojeni s ubytováním	we were satisfied with the accommodation
nachází se v centru města	is located in the city center
čisté pokoje a ochotný personál	clean rooms and helpful staff
dobře vybavený a čistý pokoj	well furnished and clean room
hotel je v klidné části	hotel is in a quiet location
dobrý poměr cena a kvalita	good quality/price ratio
velice ochotný a příjemný personál	very helpful and nice staff

Negative	
Original	**Translation**
a velmi špatná kvalita jídla	and very bad quality of food
a vysoká cena za snídani	and high price for breakfast
a hučící jednotka vnější klimatizace	and buzz of external air conditioning unit
a vysoká cena za pobyt	and high price for the stay
platba za internet za tv	payment for internet for TV
z pokoje výhled do zdi	from the room view into a wall
na jiném patře než záchod	on a different floor than WC
a pokoj byl velmi malý	and the room was small
a snídaně byly pouze sladké	and breakfasts were only sweet
v noci hluk z ulice	at night noise from the street

Table V
EXAMPLES OF DIFFERENT DISTANCES OF WORDS FORMING THE SAME SIGNIFICANT EXPRESSION "GOOD LOCATION"

Original text	Distance
...the hotel has a very good location, very...	1 word
...location quite good...	2 words
...location was very good...	3 words
...location of the hotel was very good...	6 words
...everything was good about the hotel, the staff, the location, the food...	7 words

sentences. However, the texts are not always clearly separated into sentences using correct punctuation, and also one sentence can contain two or more relatively independent parts. For example, in a phrase "It is a quiet location for a good nights sleep" the word "good" relates to "nights" and not to "location" even when these two words are just three words apart in one sentence.

- Increasing the distance from the significant word, where the search for expressions is carried out, considerably increases the computational intensity in terms of number of executed calculations as well as memory consumption. For example, increasing the length of the part of the text which is being searched from 5 to 7 words increases the number of possible word pairs from 10 to 21 and increasing the length of the part of the text which is being searched from 7 to 9 words increases the number of possible word triplets from 35 to 84. However, according to the results published by various researchers, for example see [8], there is usually no sense in generating phrases (or expressions) containing more than 4 words because longer phrases do not bring better results.

V. CONCLUSION

The goal of the presented research work was to automatically select significant expressions that would effectively characterize the positive and negative opinions of customers' reviews concerning the hotel accommodation booked via the Internet service. This service is continuously collecting its customers' opinions because it can be used for improving such a service by avoiding typical errors or utilizing the positive sides.

The reviews are in an unstructured free form, in many different natural languages. In addition, the data volumes are very big, especially for the most commonly used languages like English, Spanish, French, and German. At the present time, millions of reviews are available and processing them manually is practically impossible. This paper describes a suggested procedure how to apply computers, machine learning, and natural language processing areas to solving this described difficult problem.

The initial results were carried out with four wide-spread languages (English, German, Spanish, and others; however, the other languages are not presented here because of the limited length of the paper) and for the native language of the authors (Czech). The obtained lists of significant expressions can be considered as a good initial result because the resulting number of *significant* words was reduced from the total number 80,000–200,000 to only 200–300 and the number and quality of significant expressions is suitable for the following human processing with a sufficient quality.

Automatically generated dictionaries of the described type can be also valuable in various commercial or noncommercial areas. For example, they might be used as part of marketing research or marketing intelligence subsystems of a Marketing Information System [5], for filtering reviews, generating lists of key-words, improving parameters of search engines, and so like.

During experimenting with various large data volumes, the authors found additional possibilities and useful steps for improving the suggested procedure. This research tried to avoid the particular language dependency, however, to obtain better results, certain language properties would have to be accepted, for example, for eliminating meaningless words bringing no information to the process, improving the preprocessing phase using selected linguistic tools (removing stop-words, stemming). These goals define the plan for the intended future continuing research.

ACKNOWLEDGMENT

This research work published in this paper was supported by the Research program of Czech Ministry of Education VZ MSM 6215648904. The authors would also like to thank to the Faculty Research Center for their support.

REFERENCES

[1] M. W. Berry and J. Kogan (Editors), *Text Mining: Applications and Theory*. John Wiley and Sons, 2010.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.

[3] M. Chang and C. K. Poon, "Using phrases as features in email classification," *The Journal of Systems and Software*, vol. 82, pp. 1036-1045, 2009.

[4] F. Dařena, "Text mining-based formation of dictionaries expressing opinions in natural languages," in *Proceedings of the 17th International Conference on Soft Computing Mendel 2011*, pp 374–381.

[5] F. Dařena, "Global architecture of marketing information systems," *Agricultural Economics*, vol. 52, no. 9, pp. 432–440, 2007.

[6] D. D. Lewis, *Representation and learning in information retrieval*, Ph.D. Thesis, Amherst, MA, USA, 1992.

[7] Y. Li, S. M. Chung and J. D. Holt, "Text document clustering based on frequent word meaning sequences," *Data & Knowledge Engineering*, vol. 64, pp. 381–404, 2008.

[8] D. Mladenic and M. Grobelnik, "Word sequences as features in text-learning," in *Proceedings of the ERK-98, the Seventh Electrotechnical and Computer Science Conference*, 1998, pp. 145–148.

[9] F. Peng and X. Huang, "Machine learning for Asian language text classiffication," *Journal of Documentation,* vol. 63, no. 3, pp. 378–397 2007.

[10] <http://www.rulequest.com/see5-info.html>

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.

[12] W. Zhang, T. Yoshida and X. Tang, "Text classification based on multi-word with support vector machine," *Knowledge-Based Systems*, vol. 21, pp. 879–886, 2008.

[13] J. Žižka and F. Dařena, "Automatic Sentiment Analysis Using the Textual Pattern Content Similarity in Natural Language," Springer: *Lecture Notes in Artificial Intelligence*, vol. 6231, pp. 224–231, 2010.

[14] J. Žižka and F. Dařena, "Mining Significant Words from Customer Opinions Written in Different Natural Languages," Springer: *Lecture Notes in Artificial Intelligence*, vol. 6836, pp. 211–218, 2011.